Notes on Comparison of Bayesian Predictive Methods for Model Selection

Ryan Giordano, Nov 2019

November 7, 2019

1 Models

We're going to be talking about model selection in the linear model (eq 24)

$$\begin{aligned} y_n|x, w, \sigma^2 &\sim \mathcal{N}\left(w^T x_n, \sigma^2\right) \\ w|\sigma^2, \tau^2 &\sim \mathcal{N}\left(0, \tau^2\sigma^2 I\right) \\ \sigma^{-2} &\sim \text{Gamma}\left(\alpha_\sigma, \beta_\sigma\right) \\ \tau^{-2} &\sim \text{Gamma}\left(\alpha_\tau, \beta_\tau\right). \end{aligned}$$

For a fixed τ, this has a closed-form posterior. They authors give τ a weakly informative prior and integrate over it numerically (this is feasible because it's one-dimensional). They will also consider binary classification

$$\begin{aligned} y_n|x_n, w &\sim \text{Bernoulli}\left(\Phi\left(w^T x_n\right)\right) \\ &\text{...with the same other priors as before (note no } \sigma). \end{aligned}$$

The first entry of x_n is 1, i.e. the model always includes an intercept. The x_n are all centered and scaled to have unit variance.

Sub-models are described by γ, which is a vector of binary vectors indicating which columns of x_n are used for prediction. The prior is

$$\begin{aligned} \gamma^j|\pi &\sim \text{Bernoulli}\left(\pi\right) \\ \pi &\sim \text{Beta}\left(a, b\right). \end{aligned}$$

The values a and b are fixed, and $\gamma^0 = 1$ so that the intercept is always included.

We will denote by D the dataset (x and y) and by M the model chosen (γ). This notation allows us to discuss the selection methods in generality. But the context will always be linear regression.

2 Objectives

We are going to be looking to find

- The most parsimonious model that
- Gives good enough predictive performance.

By "parsimonious" we mean "uses the fewest variables". By "good enough" and "predictive performance" we will mean various things. Though there is obviously a tradeoff, the parameters of this tradeoff will not be articulated.

We will be considering a number of different procedures for choosing the model. See Table 2 of the paper. In my view, these procedures fall into two categories — ones that require sampling the model space and ones that do not. They also make various levels of assumption concerning the accuracy of your model.

Abbreviation	Method	Computation	Class
CV-10	10-fold CV	Forward stepwise + multi-fold	M-open
WAIC	Widely Applicable Information Criterion	Forward stepwise	M-open
DIC	Deviance Information Criterion	Forward stepwise	M-open
L2	In-sample L2 loss	Forward stepwise	M-mixed
L2-CV	Leave-one-out CV L2 loss	Forward stepwise + multi-fold	M-mixed
L2-k	L2-CV with up-weighted variance	Forward stepwise + multi-fold	M-mixed
MAP	Single most probable model	MCMC	M-closed
MPP	All variables with posterior probability > 0.5	MCMC	M-closed
BMA-ref	Smallest single model with 95% explanatory power	MCMC	M-completed
BMA-proj	Smallest projected model with 95% explanatory power	MCMC + KL projection	M-completed
BMA	Bayesian model averaging	MCMC	M-closed

2.1 What do we mean "predictive performance"?

We're going to assume that there is a true data generating distribution $p_t(\tilde{y})$ for a new datapoint $\tilde{y}$. (We always condition on x implicitly.) We will target the KL divergence between $p(\tilde{y}|D,M)$ and $p_t(\tilde{y})$ (eq 2):

$$\begin{aligned}\bar{u}(M) &= KL(p_t(\tilde{y}) \,||\, p(\tilde{y}|D,M)) \\ &= \mathbb{E}_{p_t(\tilde{y})}[\log p(\tilde{y}|D,M)] + C.\end{aligned}$$

Note that there may be no $p(\tilde{y}|D,M)$ such that this divergence is zero; in this sense, this is an M-open setting. Of course, we cannot calculate this because we don't know $p_t(\tilde{y})$. Each method amounts to a different way to estimate this intractable expectation. There may also be differences in what D and M are allowed to be.

2.1.1 Methods that do not require a reference model

What about

$$\text{Training:} \quad \bar{u}(M) \approx \frac{1}{N}\sum_n \log p(y_n|D,M)?$$

This is using the empirical distribution to replace $p_t(\tilde{y})$. Watanabe (2010) (and earlier actually, but that's the one I looked at) showed that this is biased in the sense that, even as $N \to \infty$,

$$\mathbb{E}_{p_t(\tilde{y})}\left[\frac{1}{N}\sum_n \log p\left(y_n|D,M\right)\right] \neq \mathbb{E}_{p_t(\tilde{y})}\left[\log p\left(\tilde{y}|D,M\right)\right].$$

They key problem is that this uses the observed data twice: once to form the posterior predictive $p(y_n|D,M)$, and once to approximate $p_t(\tilde{y})$. We'll see more in the WAIC.

To avoid double-using the data (term by term) we can consider (eq 3)

$$\text{CV-10:} \quad \bar{u}(M) \approx \frac{1}{N}\sum_n \log p\left(y_n|D_{-s(n)},M\right),$$

where $D_{-s(n)}$ is the data with some set $s(n)$ left out. When $|s(n)| = 1$ this is nearly unbiased, but then you have to fit N models. They use 10 folds (mostly). Still, there may be high variance.

It would be nice to approximate the CV loss without having to fit many difference models. This is the role of the "widely applicable information criterion", WAIC. Define (eq 5)

$$V := \sum_n \text{Var}_{p(\theta|D,M)}\left(\log p\left(y_n|\theta,D,M\right)\right)$$

$$\text{WAIC:} \quad \bar{u}(M) \approx \frac{1}{N}\sum_n \log p\left(y_n|D,M\right) - \frac{V}{N}.$$

This is the training set loss from above, but penalized by the average posterior variance of the log probability. Watanabe showed that WAIC is equivalent to leave-one-out CV asymptotically, but without the need to re-fit the model. In particular, he showed (Watanabe 2010, eq 6),

$$\mathbb{E}\left[\text{WAIC}\right] = \mathbb{E}_{p_t(\tilde{y})}\left[\log p\left(\tilde{y}|D,M\right)\right] + o\left(\frac{1}{N}\right).$$

Observing that $V/N = O(1)$, this also implies that the training utility is biased upwards.

The DIC seems strange and a little ad-hoc. Even reading "The deviance information criterion: 12 years on" by the original authors, it seems a bit ad-hoc, a way of coming up with a different notion of "number of variables" to plug into what is very similar to an AIC-like correction. The authors say "The success of DIC has rested largely on its ease of computation and availability in standard software." Note that DIC preceded WAIC, which might be thought of as just better, although DIC does allow for expectations of log predictives rather than of predictives, which is nice. Also Christian Robert does not like DIC. DIC may be most useful as a lesson in the relative value of implementing

something in software vs doing careful theoretical work. Anyway, here it is:

$$\begin{aligned}
\bar{\theta} &:= \mathbb{E}\left[\theta | D, M\right] \\
p_{eff} &:= 2\sum_{n=1}^{N}\left(\log p\left(y_n|\bar{\theta}, D, M\right) - \mathbb{E}_{p(\theta|D,M)}\left[\log p\left(y_n|\theta, D, M\right)\right]\right) \\
\text{DIC:} \quad \bar{u}\left(M\right) &\approx \frac{1}{N}\sum_{n}\log p\left(y_n|\bar{\theta}, D, M\right) - \frac{p_{eff}}{N}.
\end{aligned}$$

Now let's consider some approximate utilities that simple change the loss function.

$$\text{L2}: \quad \bar{u}\left(M\right) \approx -\left(\frac{1}{N}\sum_{n}\left(y_n - \mathbb{E}\left[\tilde{y}_n|x_n, D, M\right]\right)^2 + \frac{1}{N}\text{Var}\left(\tilde{y}|x_n, D, M\right)\right).$$

This is like a bias-variance decomposition for the loss $\mathbb{E}_{p_t(\tilde{y})p(y|D,M)}\left[(y-\tilde{y})^2\right]$. Apparently this is asymptotically between the training loss and CV. You can also do a cross-validated version of L2, as well as a version that up-weights the variance term for some reason. These are L2-CV and L2-k respectively.

2.1.2 Methods that require a reference model

Wouldn't it be great if we just had a good estimate of $p_t\left(\tilde{y}\right)$? Suppose we had an M_* which we are willing to accepts as a perfectly good predictive model. Then we can simply plug in

$$KL\left(p_t\left(\tilde{y}|x\right)||p\left(\tilde{y}|x, D, M\right)\right) \approx -\int p\left(\tilde{y}|x, D, M_*\right)\log p\left(\tilde{y}|x, D, M\right)d\tilde{y} + C.$$

(Remember that we get a different prediction for each x_n.) This requires an integral over the space of $\tilde{y}$, but that's typically low-dimensional. In practice, we average this over all the observed regressors.

$$\text{BMA-ref}: \quad \bar{u}\left(M\right) \approx \frac{1}{N}\sum_{n}\int p\left(\tilde{y}|x_n, D, M_*\right)\log p\left(\tilde{y}|x_n, D, M\right)d\tilde{y}.$$

Why BMA? Because they use Bayesian model averaging (BMA) over the full model as M_*. More interesting, I think, would be using some other computationally efficient but purely predictive system (regression trees, for instance) as M_* but they don't consider this for some reason. Note sampling from M_* uses MCMC. They use a reversible jump MCMC algorithm (basically MH with consideration for jumping between spaces of different dimensions).

Of course, once we have BMA for M_* we also have two more natural model guesses:

$$\begin{aligned}
\text{MAP}: &\quad M = \text{argmax}P\left(M|D\right) \\
\text{MPP}: &\quad M = \left\{\text{all regressors such that } P\left(\gamma_d|D, M_*\right) > 0.5\right\}.
\end{aligned}$$

These models do not, I think, directly maximize any version of $\bar{u}(M)$, though they make some intuitive sense. You couldn't get these naturally from a different choice of M_* like regression trees.

Note that maximizing BMA-ref amounts to minimizing the average KL divergence between M_* and M. It will be useful to give this a name:

$$\delta(M_*||M) := \frac{1}{N}\sum_n KL\left(p(\tilde{y}|x_n, D, M_*)\,||p(\tilde{y}|x_n, D, M)\right).$$

Since M_* is an acceptable model, why not choose M^*? Because maybe we actually we want a simpler model that is "not too far from M_*". More in a minute about what we mean by "not too far".

Note that, in BMA-ref, $p(\tilde{y}|x_n, D, M)$ is still formed from the posterior over the parameters θ:

$$p(\tilde{y}|x_n, D, M) = \int p(\tilde{y}|x_n, \theta, D, M)\, p(\theta|D, M)\, d\theta.$$

However, the benefit of M is its parsimony, not its posterior, as all we really care about is closeness to $p(\tilde{y}|x_n, D, M_*)$. So why not abandon the posterior and simply find the closest predictive distribution to M_* given the constraints of M? Specifically, for each θ^* in the model space of M_*, we define

$$\theta^{\perp}(\theta^*) := \text{argmin}_\theta \frac{1}{N}\sum_n KL\left(p(\tilde{y}|\theta^*, x_n, D, M_*)\,||p(\tilde{y}|\theta, x_n, D, M)\right).$$

This may seem heavy, but it turns out to be equivalent to an MLE with observations $\mathbb{E}_{p(\tilde{y}|\theta^*, x_n, D, M_*)}[\tilde{y}]$. So for simple distributions closed forms or efficient software will already exist.

Of course, despite the suggestive notation, θ^* and $\theta^{\perp}$ may live in utterly different spaces. Indeed, M_* would not even need to be representable as a marginal over some θ^*, in which case we could just minimize the KL divergence directly:

$$\theta^{\perp} := \text{argmin}_\theta \frac{1}{N}\sum_n KL\left(p(\tilde{y}|x_n, D, M_*)\,||p(\tilde{y}|\theta, x_n, D, M)\right).$$

We then have

$$\delta(M_*||M) \approx \frac{1}{N}\sum_n \mathbb{E}_{p(\theta^*|D,M^*)}\left[KL\left(p(\tilde{y}|x_n, D, M_*)\,||p\left(\tilde{y}|\theta^{\perp}(\theta^*), x_n, D, M\right)\right)\right].$$

Note that this actually is not our purported loss. In fact, it seems that what we want is

$$KL\left(\mathbb{E}_{p(\theta^*|D,M^*)}\left[p(\tilde{y}|x_n, \theta^*, D, M_*)\right]\,||\mathbb{E}_{p(\theta^*|D,M^*)}\left[p\left(\tilde{y}|\theta^{\perp}(\theta^*), x_n, D, M\right)\right]\right),$$

and in general the log and expectation don't commute. As far as I can tell, Dupuis and Robert (2003) do not comment on this discrepancy (see Section

6.1). However, in the extended survey that forms the basis of this paper, Vehtari and Ojanen (2012) do comment on the difference (page 204). This loss is called the Gibbs loss; the difference is the order of the expectation and logarithm. In this paper, only BMA-proj and DIC target the Gibbs loss. As with BMA-ref, we choose the "smallest" M with acceptable explanatory power, which we now define.

2.1.3 Explanatory power defined

We're going to choose a model that is "close" in KL divergence to the predictive model of M_*. Let's define what "close" means. Now, the absolute scale of KL divergence is not something particularly meaningful for continuous distributions, so how far is too far? Dupuis and Robert (2003) recommend specifying a scale using the "empty model":

$$\text{Explanatory power:} \quad \phi(M) := 1 - \frac{\delta(M_*||M)}{\delta(M_*||M_0)}.$$

Note that to calculate δ for this formula you need the entropy of the M_* predictive distribution; this is not invariant to additive constants. Also, the authors note that distances to a true generating model may give quite different results, so that this measure of explanatory power doesn't seem to have very good predictive performance.

A bit like a reference model, this requires some notion of "empty model" which need not be in the space of models considered, though of course it should be further from M_* than any M or the explanatory power can be negative. For regression there is, again, a natural empty model, though in general it seems like any easily-computed naive model could work.

Given this, the authors try to find

$$M = \text{Smallest } M \text{ such that } \phi(M) \geq 0.95.$$

Alternatively, we could do ALL THE THINGS. The authors describe (the end of section 2.4 and especially 4.4) the following modifications to BMA-PROJ. Rather than use $\delta(M_*||M)$ as the accuracy, do 10-fold CV to estimate the out-of sample error — that is, we use CV-10 with the models M and M_* to estimate ΔMLPD ("mean log predictive density").

$$\Delta\text{MLPD} = \frac{1}{N_{CV}} \sum_{n=1}^{N_{CV}} \left(\log\left(\tilde{y}_n|\tilde{x}_n, D, M\right) - \log\left(\tilde{y}_n|\tilde{x}_n, D, M_*\right)\right).$$

(The authors find that ΔMLPD is always positive; more on that later.) Within each fold, ΔMLPD is a random variable, and we want to choose the smallest model that is "statistically indistinguishable" from M_t. For that we need an estimate of the variability of ΔMLPD , which they propose estimating using the Bayesian bootstrap. With all this in hand, we choose the smallest model so that

$$P(\Delta\text{MLPD} \geq U) \geq \alpha,$$

for some U and α. As you can imagine, there is a computational price to be paid.

3 Results

Look at the graphs from the paper. In summary:

- BMA does best at predicting
- BMA-proj does best at reducing the model, though the other BMA-derived estimates doing reasonably well
- All the M-open methods do a fairly poor job (relatively) at doing forward stepwise regression due to their variability.

I would like to have seen some different experiments. Because the simulations were all done in M-closed, they may have been unfair to the M-open methods. Furthermore, it would be interesting to combine the M-open methods with something more sophisticated than forward stepwise regression. Similarly, it would be interesting to try the BMA-derived approaches using something less heavy than MCMC, such as draws from an approximate VB posterior. Finally, the really interesting idea seems to be combining the BMA-proj with an M_* derived from something more interesting and expressive like a regression tree or neural net.

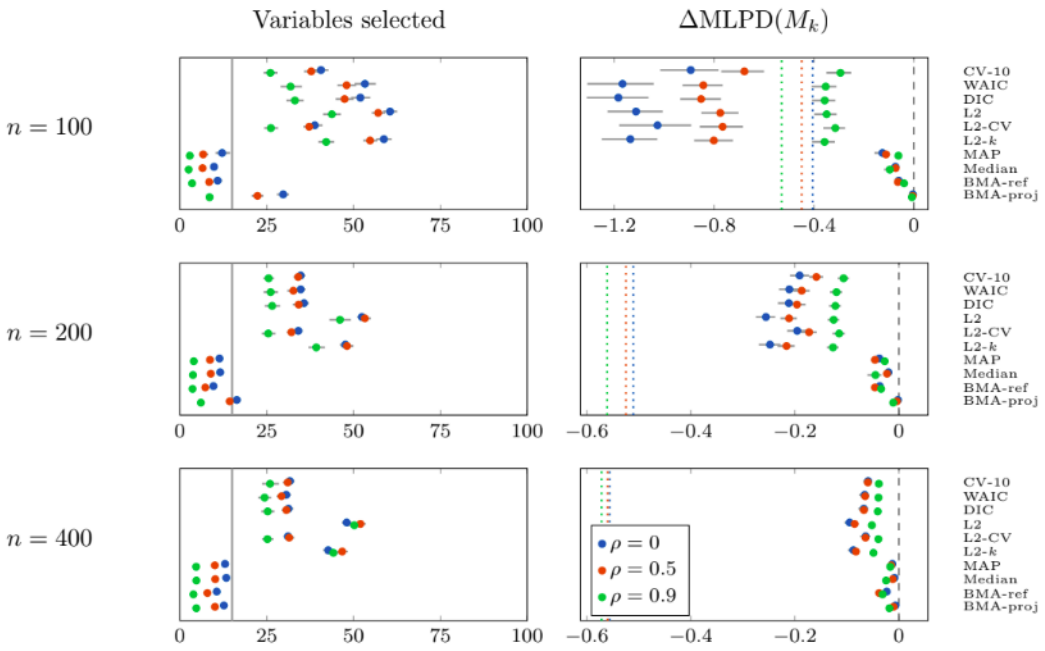

Figure 2: Simulated data: Average forward search paths for some of the selection methods for different training set sizes n when $\rho = 0.5$. Red shows the CV utility (10-fold) and black the test utility with respect to the BMA (29) after sorting the variables, as a function of number of variables selected averaged over the 50 different data realizations. The difference between these two curves illustrates the selection induced bias. The dotted vertical lines denote the average number of variables chosen with each of the methods (see Table 2).

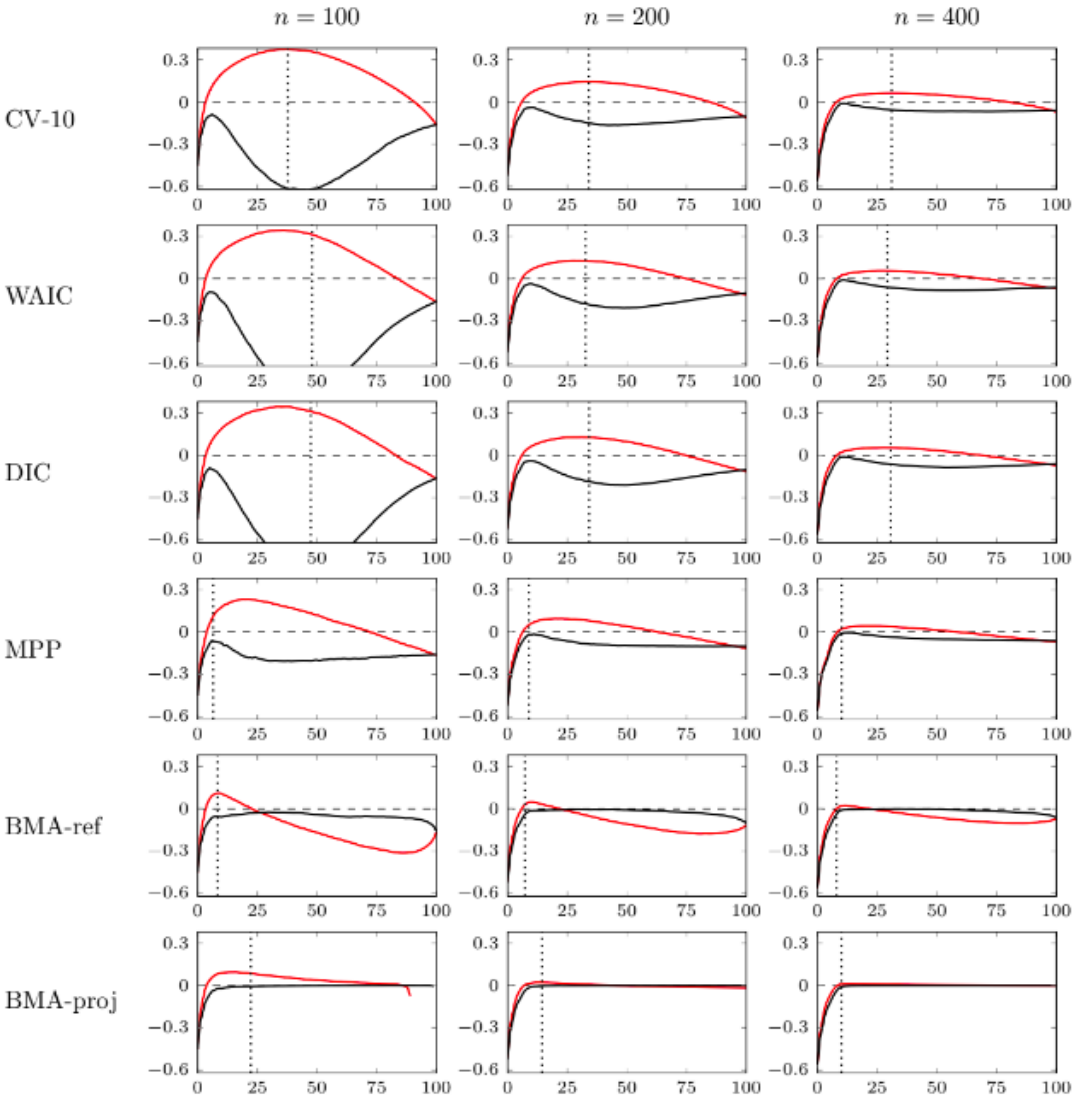

Figure 3: Simulated data: Average forward search paths for some of the selection methods for different training set sizes n when $\rho = 0.5$. Red shows the CV utility (10-fold) and black the test utility for the submodels with respect to the BMA (29) as a function of number of variables selected averaged over the 50 different data realizations. The difference between these two curves illustrates the selection induced bias. The dotted vertical lines denote the average number of variables chosen with each of the methods (see Table 2).

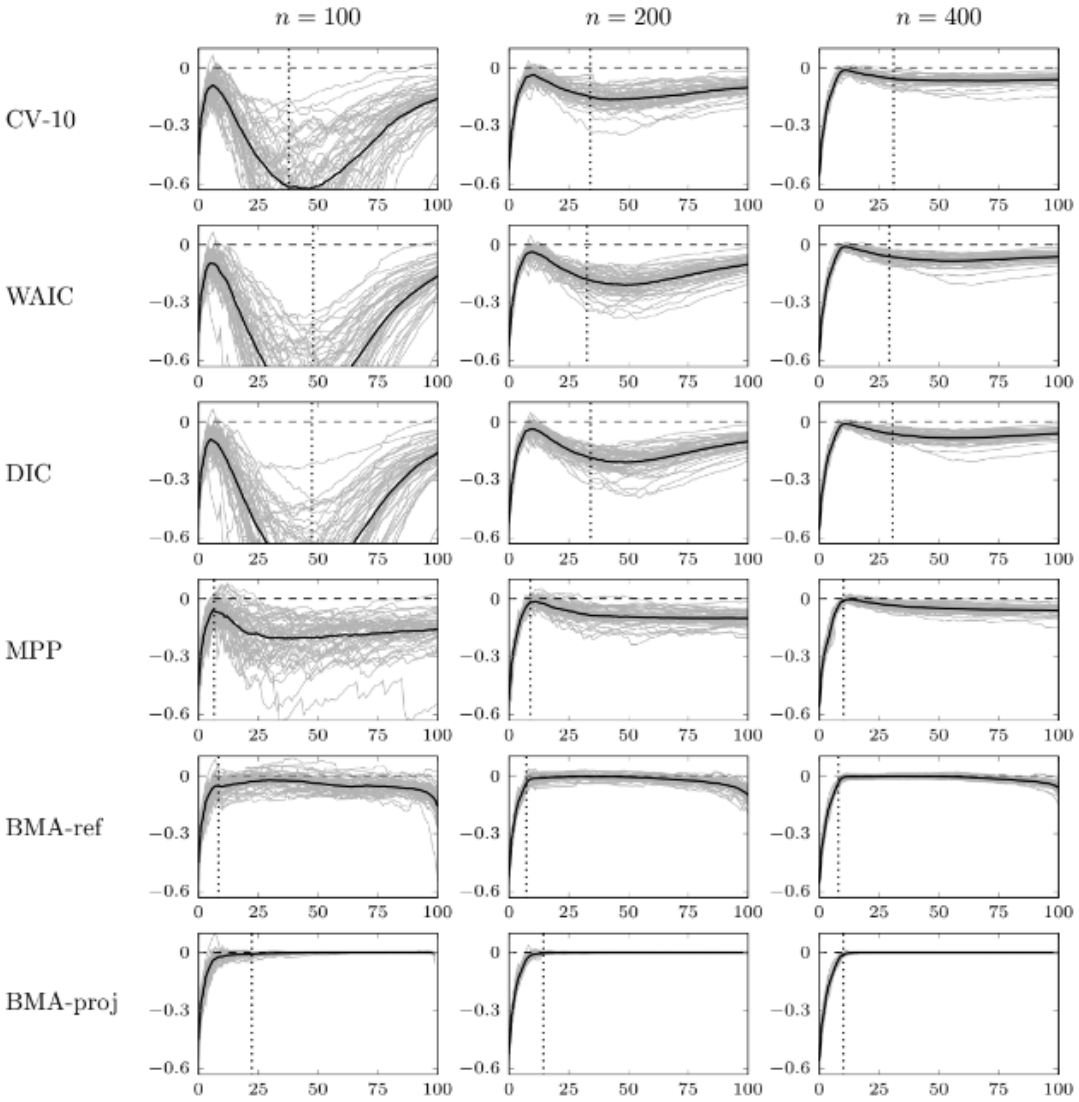

Figure 4: Simulated data: Variability in the predictive performance of the found submodels with respect to the BMA (29) along the forward search path as a function of number of variables selected for the same methods as in Figure 3 for different training set sizes n when $\rho = 0.5$. The grey lines show the test utilities for the different data realizations and the black line denotes the average (the black lines are the same as in Figure 3). The dotted vertical lines denote the average number of variables chosen.

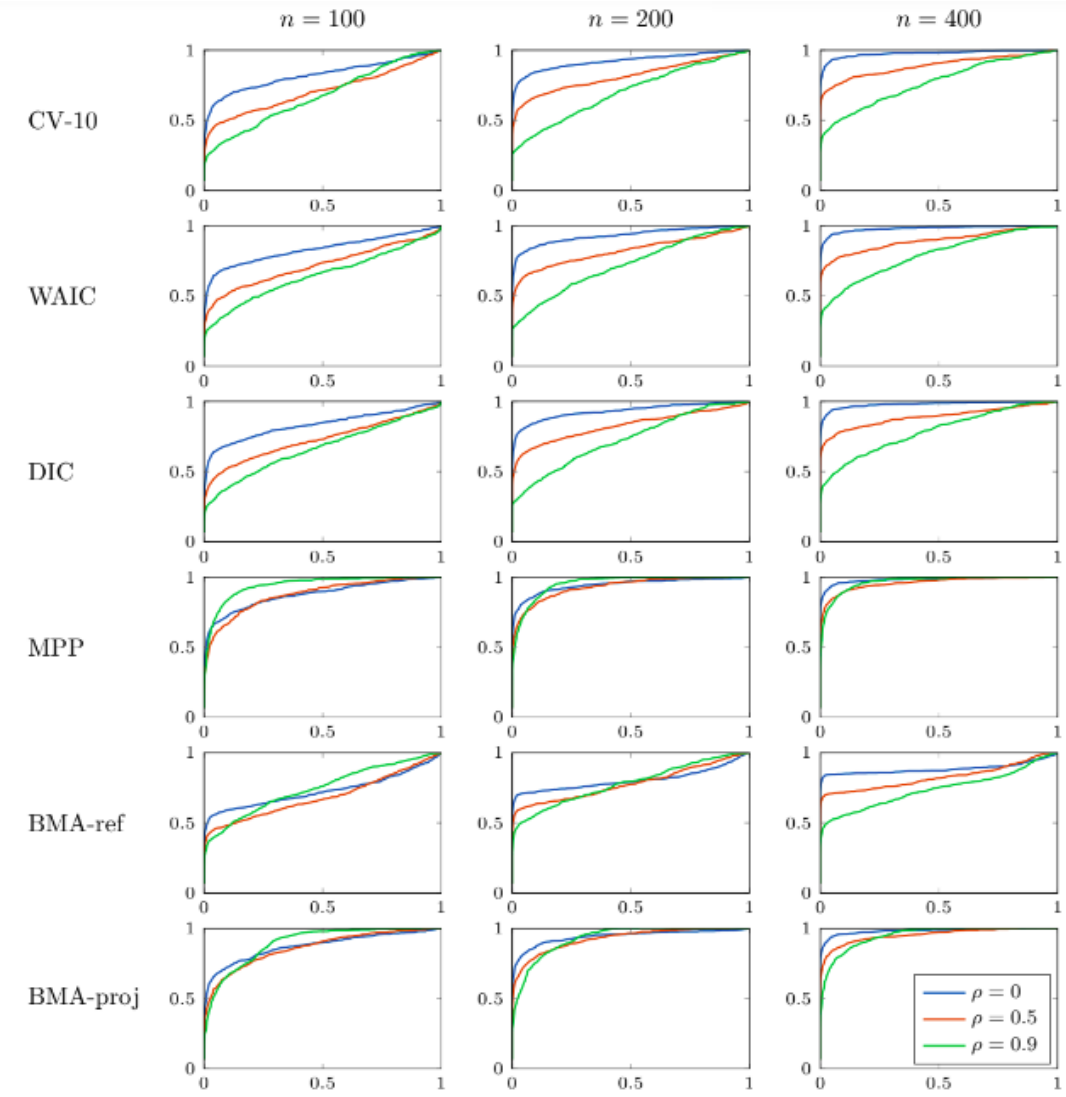

Figure 5: Simulated data: Proportion of relevant (vertical-axis) versus proportion of irrelevant variables chosen (horizontal axis) for the different training set sizes n. The data had 100 variables in total with 15 relevant and 85 irrelevant variables, relevant being defined as a variable that was used to generate the output y. The colours denote the correlation level between the variables (see the legend). The curves are averaged over the 50 data realizations.

Table 3: Summary of the real world datasets and used priors. p denotes the total number of input variables and n is the number of instances in the dataset (after removing the instances with missing values). The classification problems deal all with a binary output variable.

Dataset	Type	p	n	Prior parameters
Crime	Regression	102	1992	$\alpha_\tau = \beta_\tau = 0.5$, $\alpha_\sigma = \beta_\sigma = 0.5$, $a = b = 2$
Ionosphere	Classification	33	351	$\alpha_\tau = \beta_\tau = 0.5$, $a = b = 2$
Sonar	Classification	60	208	$\alpha_\tau = \beta_\tau = 0.5$, $a = b = 2$
Ovarian cancer	Classification	1536	54	$\alpha_\tau = \beta_\tau = 2$, $a = 1$, $b = 1200$
Colon cancer	Classification	2000	62	$\alpha_\tau = \beta_\tau = 2$, $a = 1$, $b = 2000$

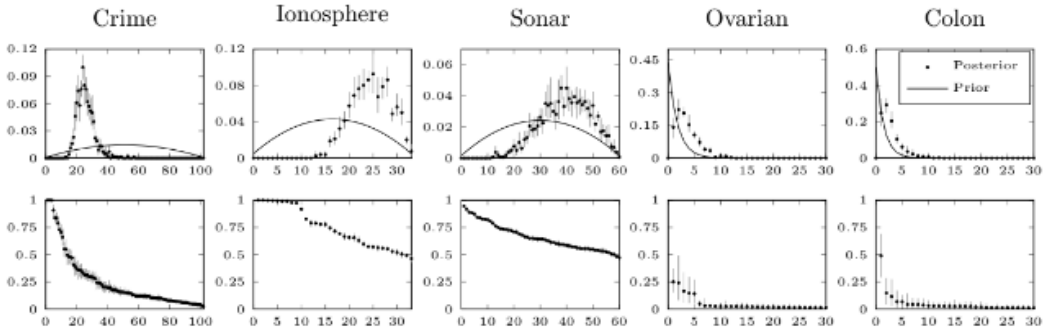

Figure 7: Real datasets: Prior and posterior probabilities for the different number of variables (top row) and marginal posterior probabilities for the different variables sorted from the most probable to the least probable (bottom row). The posterior probabilities are given with 95% credible intervals estimated from the variability between different RJMCMC chains. The results are calculated using the full datasets (not leaving any data out for testing). For Ovarian and Colon datasets the plots are truncated at 30 variables.

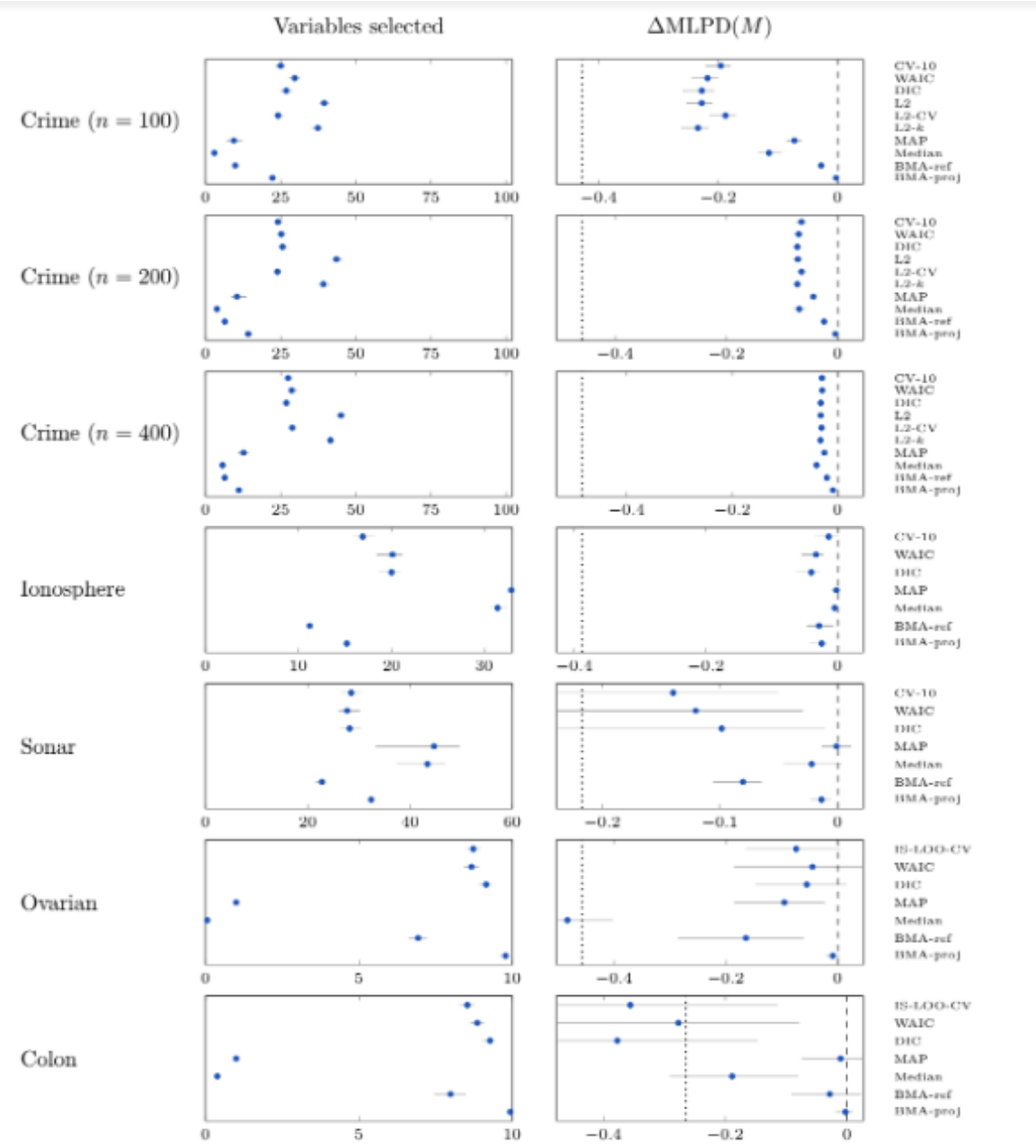

Figure 8: Real datasets: The number of selected variables (left column) and the estimated out-of-sample utilities of the selected models (right column) on average and with 95% credible intervals for the different datasets. The out-of-sample utilities are estimated using independent data not used for selection (see text) and are shown with respect to the BMA (29). The dotted line denotes the performance of the empty model (the intercept term only). For Ovarian and Colon datasets the searching was performed only up to 10 variables although both of these datasets contain many more variables.

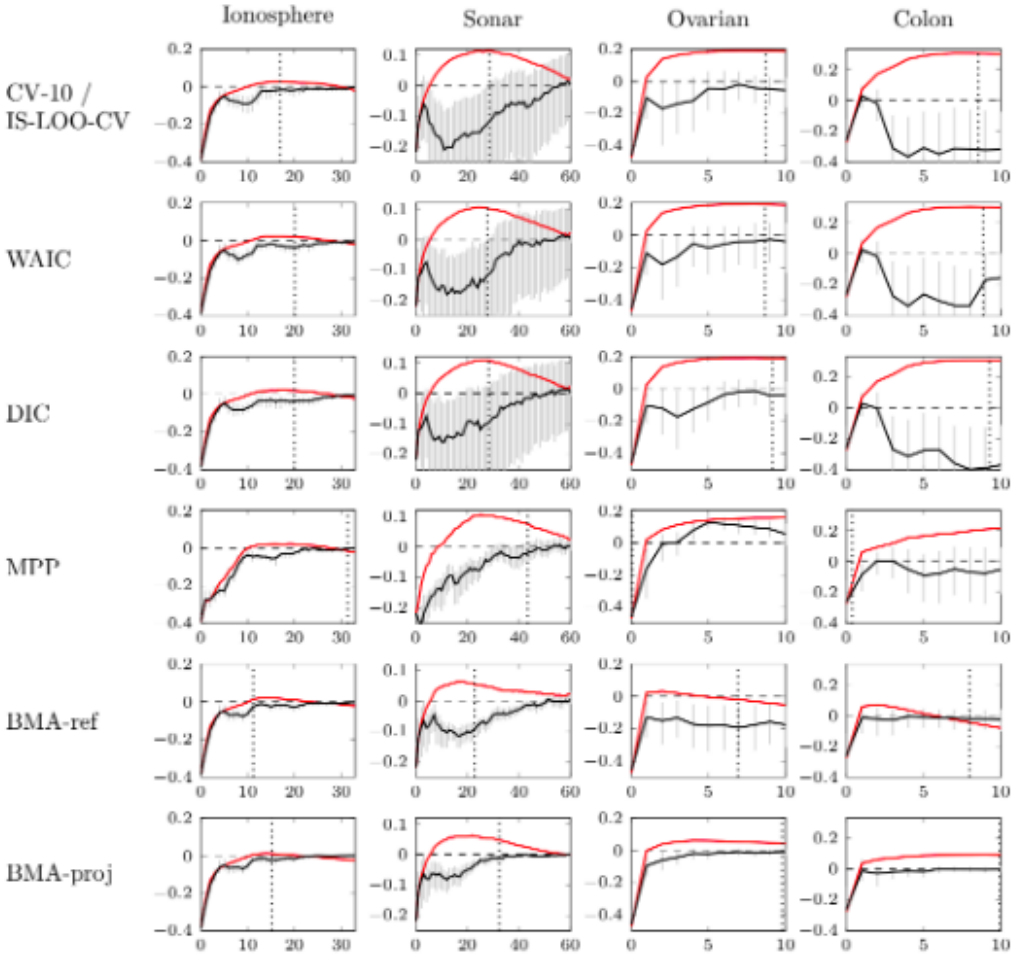

Figure 9: Classification datasets: CV (red) and out-of-sample (black) utilities on average for the selected submodels with respect to the BMA (29) along the forward search path as a function of number of variables selected. CV utilities (10-fold) are computed within the same data used for selection and the out-of-sample utilities are estimated on hold-out samples not used for selection (see text) and are given with 95% credible intervals. The dotted vertical lines denote the average number of variables chosen. CV optimization (top row) is carried out using 10-fold-CV for Ionosphere and Sonar, and IS-LOO-CV for Ovarian and Colon.

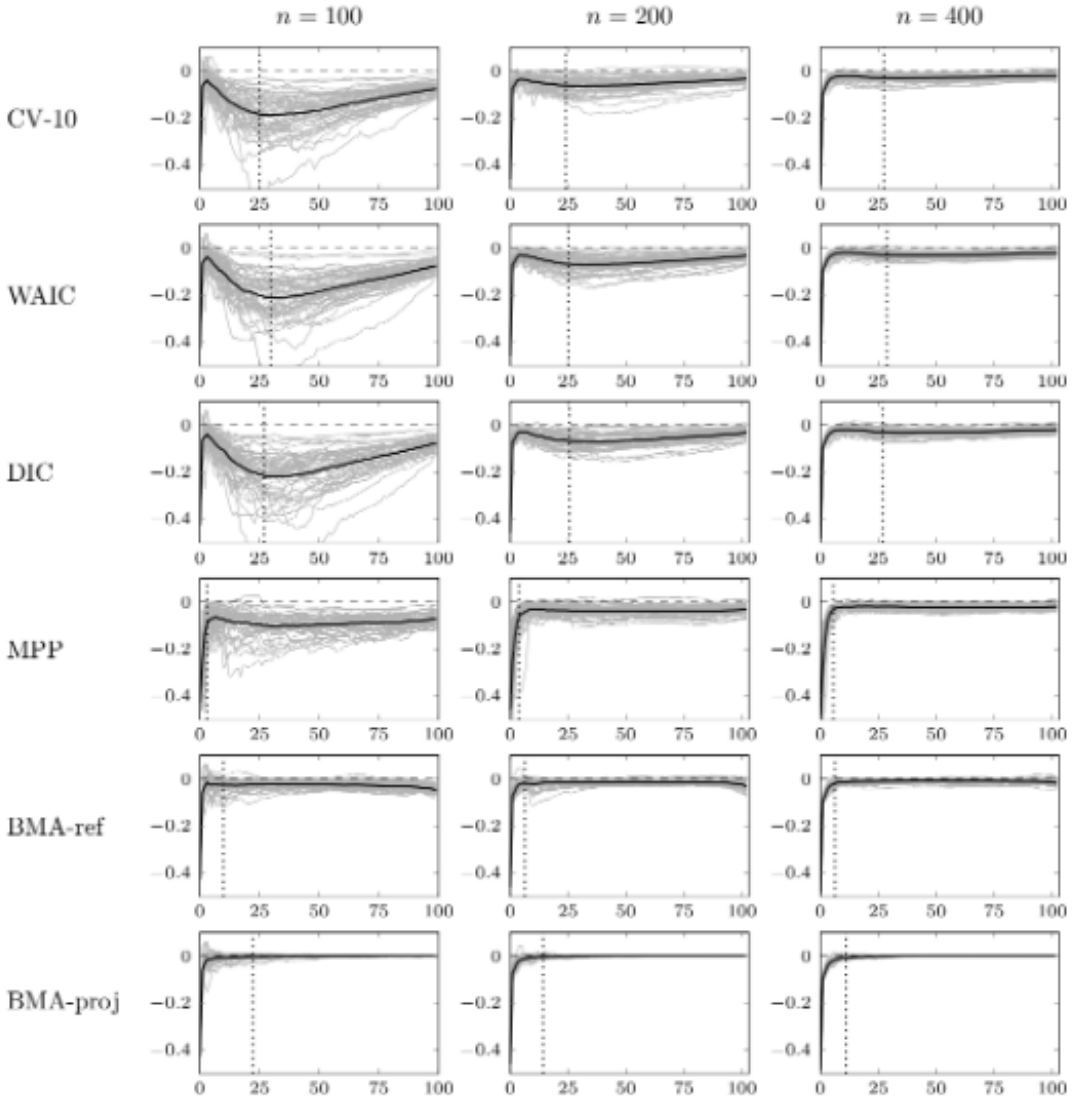

Figure 10: Crime dataset: Variability in the test utility of the selected submodels with respect to the BMA (29) along the forward search path as a function of number of variables selected. The selection is performed using $n = 100, 200, 400$ points and the test utility is computed using the remaining data. The grey lines show the test utilities for the 50 different splits into training and test sets and the black line denotes the average. The dotted vertical lines denote the average number of variables chosen.

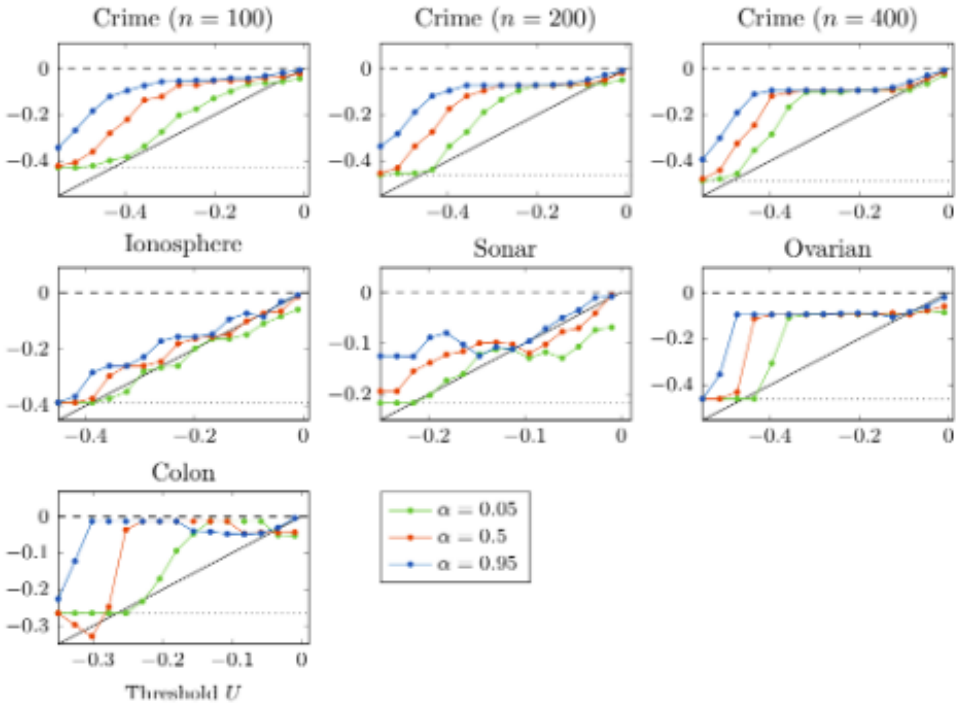

Figure 11: Real datasets: Vertical axis shows the final expected utility on independent data with respect to the BMA (29) for the selected submodels when the searching is done using the projection (BMA-proj) selecting the smallest number of variables m satisfying $\Pr\left[\Delta\text{MLPD}(m) \geq U\right] \geq \alpha$, where $\Delta\text{MLPD}(m)$ denotes the estimated out-of-sample utility for m variables estimated using the CV (10-fold) outside the searching process (same as the black lines in Figure 9). The final utility is estimated using another layer of validation (see text). The dotted line denotes the utility for the empty model. When $\alpha = 0.95$, the final utility remains equal or larger than U (the dots stay above the diagonal line) indicating that the applied selection rule does not induce bias in the performance evaluation for the finally selected model.